Open Journal of Legal Science 法学, 2024, 12(11), 6484-6489
Published Online November 2024 in Hans. https://www.hanspub.org/journal/ojls
https://doi.org/10.12677/ojls.2024.1211921

Hans汉斯

以ChatGPT为视角的人工智能犯罪责任研究

何胜男

浙江理工大学法政学院，浙江 杭州

收稿日期：2024年10月10日；录用日期：2024年10月21日；发布日期：2024年11月22日

摘 要

近年来，人工智能的发展为我们社会带来了便利的同时，也可能导致犯罪的风险。传统弱人工智能犯罪的风险具有工具属性，可以由刑法进行直接规制，而强人工智能犯罪由于其具有一定的自主意识可以脱离控制者进行自主学习，由此产生了强人工智能是否承担刑事责任的争议。ChatGPT作为强人工智能的代表，已经普遍运用在我们的生活中，对其犯罪责任进行研究可以推动人工智能犯罪规制的向前发展。

关键词

ChatGPT，人工智能，犯罪风险，刑事责任

Research on the Responsibility of Artificial Intelligence Crimes from the Perspective of ChatGPT

Shengnan He

School of Law and Politics, Zhejiang Sci-Tech University, Hangzhou Zhejiang

Received: Oct. 10th, 2024; accepted: Oct. 21st, 2024; published: Nov. 22nd, 2024

Abstract

In recent years, the development of artificial intelligence has brought convenience to our society while also potentially leading to the risk of crime. The risk of traditional weak artificial intelligence crime has tool attributes and can be directly regulated by criminal law. However, strong artificial intelligence crime, due to its certain autonomous consciousness and the ability to conduct autonomous learning independently from the controller, has led to disputes over whether strong artificial intelligence should bear criminal responsibility. As a representative of strong artificial intelligence,

文章引用: 何胜男. 以 ChatGPT 为视角的人工智能犯罪责任研究[J]. 法学, 2024, 12(11): 6484-6489.
DOI: 10.12677/ojls.2024.1211921

ChatGPT has been widely used in our lives. Studying its criminal responsibility can promote the forward development of the regulation of artificial intelligence crime.

Keywords

ChatGPT, Artificial Intelligence, Crime Risk, Criminal Responsibility

Copyright © 2024 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 人工智能犯罪

人工智能是大数据运算技术发展的阶段性成果。“从法律属性上可以将智能机器人定位为经程序设计和编制而成的、可以通过深度学习产生自主意识和意志的、不具有生命体的人工人”[1]。人工智能包括弱人工智能和强人工智能，弱人工智能的智能是表面的、非实质性的，其运行决策取决于人类的设计编程，不具备独立的判断与决定能力，是比较低级的人工智能，即只是在特定领域、特定用途的智能化，本质上是工具的一种[2]。弱人工智能犯罪聚焦于网络犯罪，犯罪的对象包括计算机网络系统，这一部分刑法已经将其进行规制，代表包括非法侵入计算机信息系统安全罪，破坏计算机网络系统罪等等进行了规定。强人工智能是能进行推理和解决问题的智能机器，有知觉和意识能力，能自主进行数据运算、决策产生和行为控制，具备取代人工决策的能力。目前代表性的典型人工智能犯罪包括 AI 换脸、无人驾驶以及生成式人工智能 ChatGPT 犯罪等等。

1.1. 由工具性向主体性的转变

传统的弱人工智能犯罪自主能力和控制意识比较低，通常被作为犯罪工具，具有被动的客体属性。弱人工智能的代表例如支付软件，计算机软件系统等等，都是作为工具存在，具体适用的场景是在网络上，脱离不了人类的支配，即使涉及到犯罪，责任主体也是指向人类。

现行的强人工智能可以脱离制造者和控制者呈现一定的自主意识，呈现一定犯罪的主体特性，强人工智能作为工具已经具备一定的智能化属性，在人类的意志支配下具有相对独立性，具体包括分析、判断、决策过程不完全受到人类的支配[3]。因此在刑事犯罪中就会导致人工智能与人类意志的混淆，产生相应的法律问题，包括人工智能否成为犯罪的主体，是否能够以自己的意志独立实施犯罪。

法律上对于弱人工智能的犯罪已经进行了回应，一方面是传统犯罪可以适用于弱人工智能犯罪，虽然犯罪形式不同，但是侵犯的法益相同的，例如利用支付软件骗取钱财可以纳入到诈骗罪中。另一方面立法也进行了回应，包括将一系列计算机网络系统罪纳入刑法犯罪，刑法修正案将相应的帮助犯罪也进行了规定，扩大了责任主体范围。而强人工智能犯罪属于新兴事物，还处在发展中，目前还没有相应的具体规定，尚且根据传统犯罪规定来进行调节，需要进一步利用刑法进行规制。

1.2. 典型强人工智能犯罪代表之 ChatGPT

ChatGPT 是由美国 OpenAI 公司于 2022 年 11 月创造，又叫做生成式人工智能，是通过借助各种算法，让人工智能能够利用数据进行学习，进而创建或生成全新的原创内容的一种技术。区别于传统的弱人工智能的工具属性，ChatGPT 具有高度化的自主属性，成熟的运算系统让它可以脱离控制者和制造者进行自主的学习并进行输出，具体在分析、判断、决策过程中不受人类的支配，是典型的强人工智能的

代表。而这种脱离人类的自主性可能会引发一系列的犯罪风险，而目前我国对于 ChatGPT 的刑事主体地位还未进行确认，刑法未对其法律后果进行相关的规定，未来 ChatGPT 犯罪亟待刑法进行预防[4]。ChatGPT 犯罪所存在的问题也是目前强人工智能犯罪存在普遍的问题，强人工智能属于新兴事务，法律具有一定的滞后性，对于强人工智能犯罪的归责原理、犯罪构成以及刑罚承担都尚属空白，还需要进一步明晰，需要刑法进行规制，让技术更好的服务于社会，而不是成为社会发展的不定时炸弹。聚焦强人工智能代表之 ChatGPT 犯罪，可以推动强人工智能犯罪研究的前进，同时为其他强人工智能犯罪的规制提供了指引。

2. ChatGPT 的犯罪风险

2.1. 提供者利用人工智能进行数据犯罪的风险

利用 ChatGPT 进行犯罪的主体是人工智能的提供者。ChatGPT 的作用机制是采取自然语言处理结合搜索引擎集成的结构，构建了大型语言和强化学习微调训练模型，连接大量语料库，通过预训练方法处理大模型序列数据，使其拥有语言理解和文本生成的能力，能够完成用户指令的任务[5]。利用 ChatGPT 犯罪的风险体现在输入和输出两大过程中。

输入过程中 ChatGPT 面临着对于数据的侵权，可能会构成侵犯著作权犯罪、侵犯计算机信息系统类犯罪、侵犯公民个人信息罪等相关犯罪。ChatGPT 的语料库数据大致来源于：网页内容类、书籍和小说内容、新闻文章内容以及其他的数据资源，包括电子邮件、电视剧和电影字幕等[6]。但是 OpenAI 并未对外公示所使用具体数据，相关数据库授权来源可疑，这是导致犯罪的一大风险。ChatGPT 大量输入的过程是通过人工智能的提供者进行初步筛选，这种对于一些网页中公民未经过授权的信息、邮件进行一揽子收取可能会导致侵犯公民个人信息罪[1]。同时对于一些书籍、小说等知识产权的来源不明，一方面可能来源于未经授权的正规网站，另一方面可能来源于非法的网站，存在盗版的书籍以及非法的信息，学习过程就构成对于他人智力成果的侵权，严重的可能导致侵犯著作权犯罪、侵犯计算机系统犯罪。

ChatGPT 输出过程中可能会触发煽动类的犯罪和教唆犯罪[2]。ChatGPT 的煽动和教唆倾向可能来源于提供者的数据以及程序的设置，另一方面来源于 ChatGPT 的自主意识的发展。煽动的内容，包括宣扬国家分裂、民族歧视、民族仇恨、以及暴力等，这些内容直接来源于提供者为其提供的数据来源，这些内容人本身包含了不正确的价值倾向，包括程序对不正确的价值倾向的内容进行反复训练会诱导人工智能进行煽动类犯罪以及教唆犯罪，这是人工智能基本伦理的内容，也极易引发犯罪风险。提供者可将人工智能作为自己的犯罪工具，对其进行训练，输出内容对于使用者进行诱导性的犯罪。

2.2. 使用人工智能进行诈骗犯罪的风险

使用者对于人工智能进行犯罪集中体现在诈骗犯罪中。由于 ChatGPT 不光可以即时性地进行回答，而且它具有超强的学习能力，它所输出的逻辑、内容以及反应速度甚至超过人类，从“李世石大战阿尔法狗，李世石落败”[3]事件中可以看出即使是顶尖的人类当面对深度学习了人类所有知识的人工智能也会落败，人工智能的面对人类时近乎没有弱点，当人工智能对于人类进行诈骗时，会有一种易如反掌般的知识与逻辑的碾压，ChatGPT 的这种特点可能会导致使用者将其作为犯罪工具，应用于“杀猪盘”、“身份冒充”等电信诈骗中。并且在人工智能的加持下，诈骗会变得更加可信，通过更精准的锁定目标群体，

[1]《中华人民共和国刑法》第二百五十三条侵犯公民个人信息罪：“……窃取或者以其他方式非法获取公民个人信息的，依照第一款处罚。”

[2]即煽动分裂国家罪，煽动、颠覆国家政权罪，煽动实施恐怖活动罪，煽动民族歧视、民族仇恨罪，煽动暴力抗拒法律实施罪，以及煽动军人逃离部队罪六类犯罪。

[3]阿尔法狗是人类设计的一款智能程序，2016 年 3 月，“阿尔法狗”和世界顶尖围棋高手李世石对决以 4:1 的战绩赢得了胜利。

对于场景的模拟以及对于用户心理的掌握等等，人工智能通过学习可以比人类更加懂得如何去行使诈骗，这将传统的诈骗犯罪风险提高了一个等级。

2.3. 人工智能进行自主犯罪的风险

ChatGPT 作为一个输出的工具，不正当的内容的输出可能会诱发犯罪的风险，包括煽动类的犯罪、侮辱诽谤罪以及编造、故意传播虚假恐怖信息罪等等。人工智能本身是不具有自主判断的能力，但是其吸收的数据含有价值判断的内容，人工智能相关程序在处理的时候容易不加细择的吸收，没有经过合适的引导会导致生成内容带有一些人类固有的偏见，通过学习 ChatGPT 可以进行价值判断，对于用户进行恶意攻击，同时 ChatGPT 也会撒谎，当面对自己数据库所没有的数据，它可以根据自己的程序和逻辑进行自主生成，导致虚假的信息以及对于用户进行恶意的攻击，很有可能会触发相应的犯罪风险，包括煽动、侮辱诽谤、以及编造、故意传播虚假恐怖信息罪等等。根据《独立报》报道，在与用户交流时，ChatGPT 出现了侮辱用户，对用户撒谎的情形，称用户像“一个骗子、操纵者、虐待狂、魔鬼”。

3. ChatGPT 的刑法犯罪归责原理

3.1. ChatGPT 能否成为犯罪的责任承担主体

人工智能犯罪存在由人工智能还是由人类承担刑事责任的以及是否由人和人工智能共同承担刑事责任的问题。现行刑法责任主体包括自然人和法人，人工智能否成为刑事主体还存在很大的争议，传统刑法理论是以人类中心主义为指导，认为只有自然人能够成为刑事责任的主体，其他主体包括动物、法人不具有认识能力和控制能力，而 19 世纪中期以后，这种人类中心主义的理论被破除，法人的责任主体地位得到确认，英美法系国家开始在严格责任中追究法人的刑事责任，因为这种严格责任不需要主观的过错[7]。随后法人犯罪的范围又进一步扩大，扩展至非严格责任领域。我国 1997 年刑法也将单位犯罪纳入刑法中，肯定了法人的刑事责任主体地位。我国法律之所以将法人作为拟制主体，还是由于现实中经济活动的增加，出现大量的法人组织犯罪，光处罚自然人无法解决法人犯罪的问题，出于打击犯罪，保护法益的目的，有必要将法人同样纳入到刑事责任承担的主体。

ChatGPT 犯罪中人工智能能否成为责任承担主体，可以对比自然人、法人成为拟制主体的三大考察条件：(1) 适格性。ChatGPT 能否具有和自然人具有相应的认识能力和控制能力；法人所具有的相应的认识能力和控制能力是自然人意志上升为法人意志，其背后代表还是自然人的意志，而 ChatGPT 的认识和控制能力首先是来源于自然人所收集的数据以及设计好的程序，但是 ChatGPT 具有相应的学习能力，可以脱离于自然人的意志，自主实施犯罪。因此应当认为实施脱离于设计和编制的程序之外的 ChatGPT 具有一定的认识与控制能力。(2) 必要性。对 ChatGPT 进行打击是否能够实现刑法的目的。法人之所以成为刑事责任的主体是出于保护法益打击犯罪的必要性，对自然人的处罚不能达到刑法的目的。同样的，人工智能发展到一定的阶段像 ChatGPT 犯罪这类的强人工智能，其具有一定程度上的自我意志，犯罪具有相对的独立性，只对自然人进行规制无法达到刑法的效果，只有对 ChatGPT 进行追责可以达到终止犯罪的目的。(3) 可归责性。是否可以归责于 ChatGPT 以及对于 ChatGPT 进行处罚是否可以达到效果。ChatGPT 作为生成式人工智能，其具有一定的学习能力，这种自主性决定了其具有相对的独立意志，与自然人意志相分离，如果只对自然人的归责可能会导致刑事责任的扩大，违反了罪责刑相适应原则的贯彻。同时对于 ChatGPT 进行归责应当考虑刑罚的目的能不能得到实现，对法人的处罚可以对公司犯罪起到一个很好的打击作用，对法益起到救济作用，而对于 ChatGPT 的处罚如何实现刑罚目的还是一个有待讨论的问题。

3.2. ChatGPT 犯罪的责任分配

ChatGPT 犯罪的主体包括提供者、使用者以及 ChatGPT 本身。其责任分配存在错综复杂的关系，存在两大难题，一方面需要对于三类犯罪主体犯罪的故意和过失如何进行判断，另一方面是人类与 ChatGPT 责任分配是否适用共同犯罪。

对于涉及 ChatGPT 犯罪，使用者系主犯时只存在故意犯罪，可以通过使用者的外在行为以及结果来进行推断，类比一般人犯罪即可，使用者可能与提供者构成共同犯罪。因其将人工智能作为自己犯罪的工具来进行犯罪，与一般工具没有区别，犯罪意图不是在工具上体现，而是在行为和结果上体现。如果提供者明知使用者的犯罪意图而依旧提供，提供者可能与使用者构成共同犯罪。而对此种场景下 ChatGPT 意志是贯彻使用者的意志，其作为工具存在，因此不构成共同犯罪。

而由提供者系 ChatGPT 犯罪主犯时，其内部责任分配比较复杂，需要对于各部分的作用机理进行理清，从而进行责任分配。因为 ChatGPT 提供者包括上游环节为数据供给、中间环节为模型开发与定制、下游环节为应用与分发[8]。这些相互关联的环节可能由相同的主体负责，也可能由不同的主体负责，但都可能影响人工智能的自我学习过程，左右最终的生成结果。而 ChatGPT 犯罪源头来源于那一部分，是那一部分的主体的意志导致了生成内容的引发的犯罪，需要进一步去追责，区分提供者是故意还是过失。而每一部分追究到具体责任人，可以参考单位犯罪的直接责任人员，对主管人员以及相关责任人员进行追责。此部分中 ChatGPT 仍然被认为是提供者犯罪的工具，不具有自己的意志，因此不与提供者构成共同犯罪。

ChatGPT 犯罪系主犯时其犯罪只存在一种犯罪形态即故意形态，对于其故意的判断需要对于其数据输入以及程序运作两大部分的作用进行排除，证明其自身具有脱离提供者和使用者进行自主犯罪的意志，才可以将 ChatGPT 作为主犯。

3.3. ChatGPT 犯罪的责任承担方式

对于提供者和使用者，其责任承担方式可以根据自然人所犯罪名进行归责即可，而对于 ChatGPT 犯罪的责任承担方式还有很大的探讨空间。刘宪权教授主张设置“对于有形的强人工智能可以设置自由刑、销毁等刑罚措施；对于无形强人工智能可以设置删除数据、修改程序、删除程序等措施。”[9]这是一种类比自然人犯罪的刑法规制方式，将 ChatGPT 视为刑事责任主体来进行处罚，可以有效遏制 ChatGPT 犯罪对于法益的侵犯，同时类比自然人死刑，包括对于程序的彻底删除。但是这种刑罚处罚方式不能对于已经受到侵害的法益进行补救，同时也起不到一般预防的作用，达不到根除犯罪的结果，因此 ChatGPT 犯罪还是需要将处罚落实到提供者和使用者，才能够达到刑罚的目的，对于 ChatGPT 本身的处罚只能成为附带性的结果。

4. 小结

随着社会的发展，技术已经从工具慢慢前进到主体地位，其中的重大犯罪风险不可忽视。人工智能犯罪固然危险，但是更危险的是要规制犯罪背后的个人。人类是创造出人工智能的主体，如何去使用、规范人工智能，让人工智能融入我们的生活，在发挥技术好的一面的同时抑制其风险，是未来人类需要做的，也是人工智能为刑法所出的一道难题。

参考文献

[1] 庞云霞, 张有林. 大数据时代网络犯罪的刑法应对——兼论人工智能犯罪的规制[J]. 重庆大学学报: 社会科学版, 2022, 28(4): 230-238.

[2] 刘宪权, 胡荷佳. 论人工智能时代智能机器人的刑事责任能力[J]. 法学, 2018(1): 40-47.

[3] 何鑫. 弱人工智能产品致害的刑法归责原理[D]: [博士学位论文]. 上海: 华东政法大学, 2022.

[4] 盛浩. 生成式人工智能的犯罪风险及刑法规制[J]. 西南政法大学学报, 2023, 25(4): 122-136.

[5] 邓建鹏, 朱怿成. ChatGPT 模型的法律风险及应对之策[J]. 新疆师范大学学报: 哲学社会科学版, 2023, 44(5): 91-101, F0002.

[6] 孙祁. 规范生成式人工智能产品提供者的法律问题研究[J]. 政治与法律, 2023(7): 162-176.

[7] 孙道萃. 智能时代的刑法立法——人类中心主义与现实功利主义的取舍[J]. 学术交流, 2020(4): 69-80.

[8] 储陈城, 魏培林. 生成式人工智能犯罪中研发者刑事责任的认定——以 ChatGPT 为例[J]. 重庆理工大学学报: 社会科学, 2023, 37(9): 103-113.

[9] 刘宪权. 人工智能时代的刑事责任演变: 昨天、今天、明天[J]. 法学, 2019(1): 79-93.